Multivariate Regression

Team: 3

1. Shivanshu
2. Abhishek Kumar
3. I Sai Pradeep
4. Hari Vamshi

Glimpse about the Dataset

Data Description

The dataset provided for this regression challenge encompasses various socio-economic and healthcare indicators aggregated from multiple sources including the American Community Survey, clinicaltrials.gov, and cancer.gov. The primary objective of this dataset is to explore the relationship between these factors and the target variable, "TARGET_deathRate," which represents the mean per capita (100,000) cancer mortalities.

Data Sources

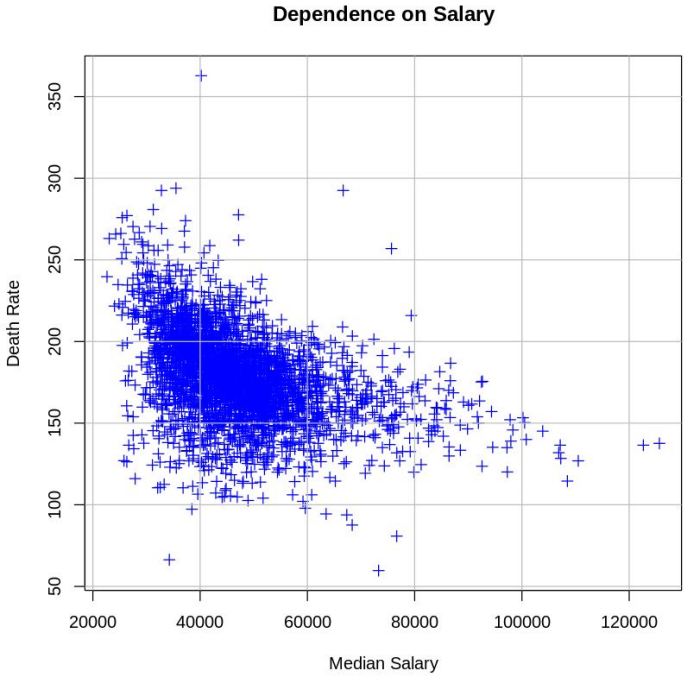
The dataset is derived from reliable sources including the American Community Survey, clinicaltrials.gov, and cancer.gov, ensuring the credibility and relevance of the collected information. The dataset can be found here

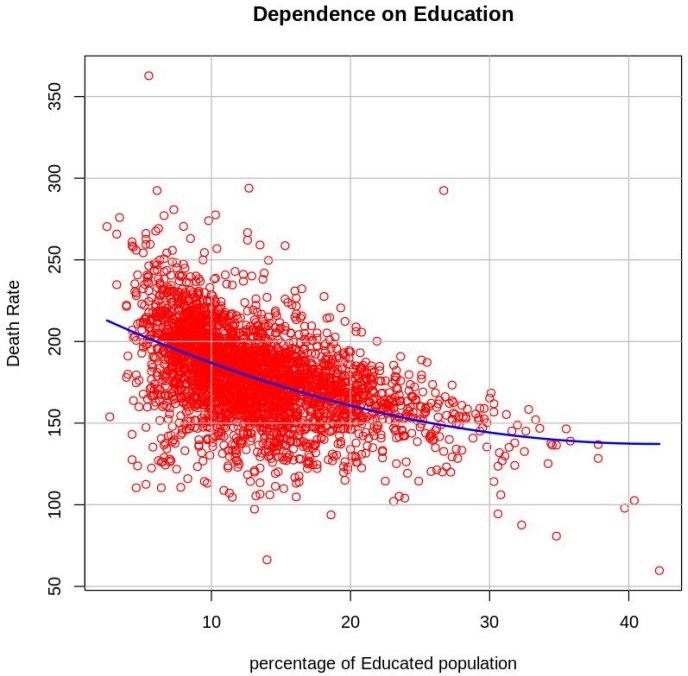
Exploratory Data Analysis(EDA)

1. Data Visualization
2. Factor Analysis(FA)
3. Clustering Analysis

Data visualization

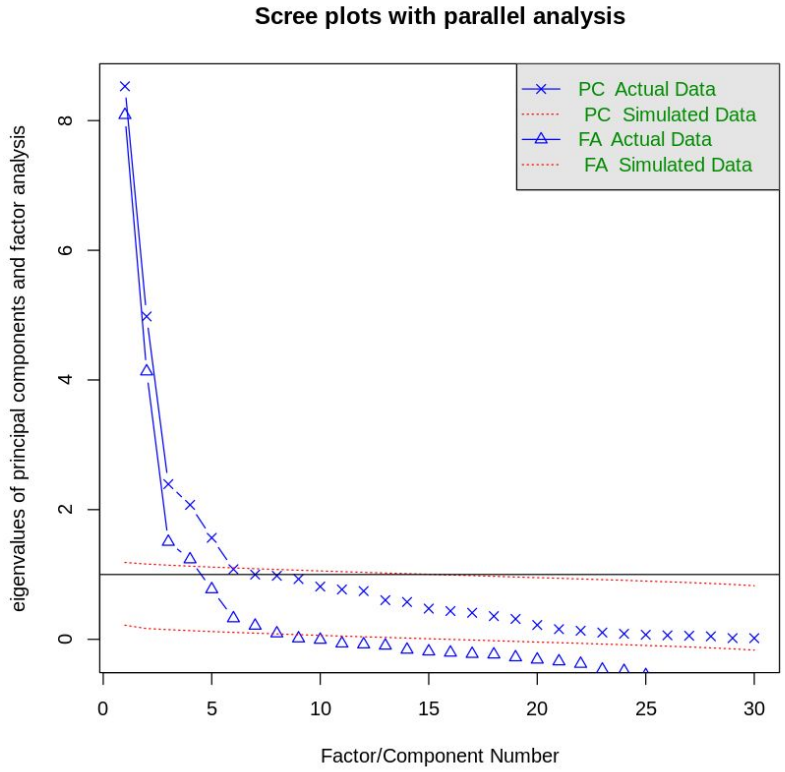
Some scatter-plots of dependent and independent variables





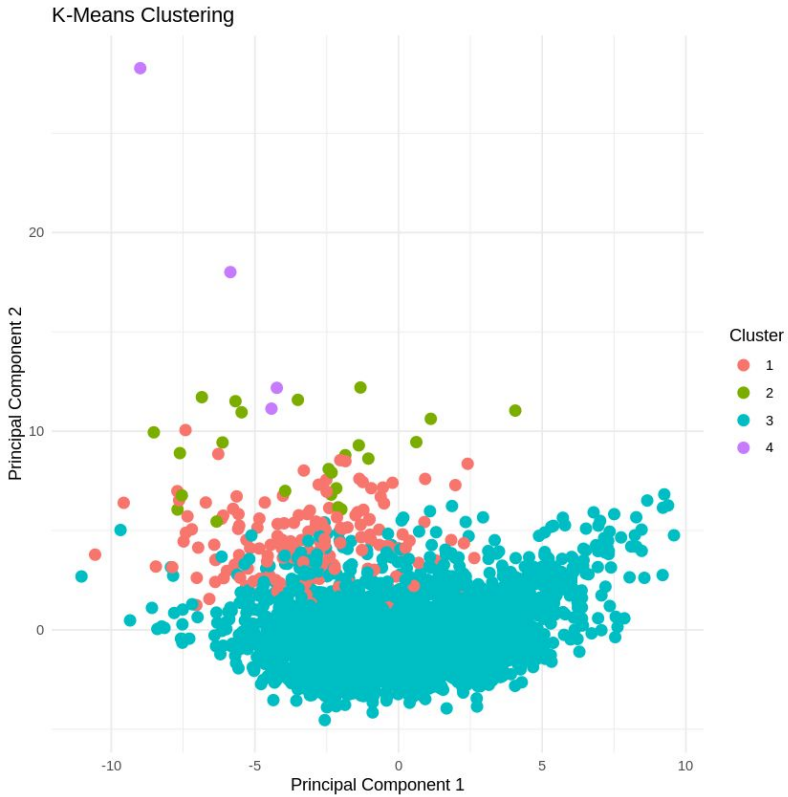
Factor Analysis

- We perform Factor Analysis to explore the underlying structure of our dataset.
- Parallel analysis suggested retaining 7 factors and 5 components based on eigenvalues obtained from random data.
- It allowed us to focus on the variables with high loadings (> 0.5) for each factor. These variables likely to explain most of the variance of the data.



K-means Clustering

- Since our data has multiple dimensions, we used PCA for dimensionality reduction to visualize the clusters in 2D space.
- After using PCA we used first two primary component as x and y for for plotting.
- For finding optimal no. of clusters(K) we used elbow method in WSS plot and from there we got 4 clusters.



Assumptions of Regression

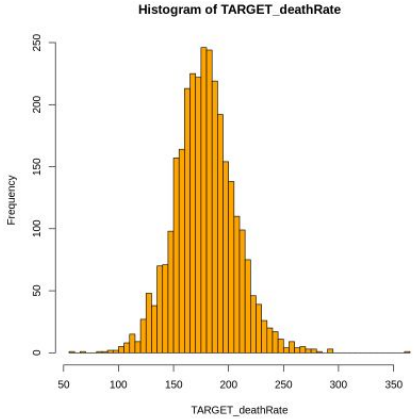
1. Zero mean
2. Normality
3. Heteroskedasticity
4. Collinearity

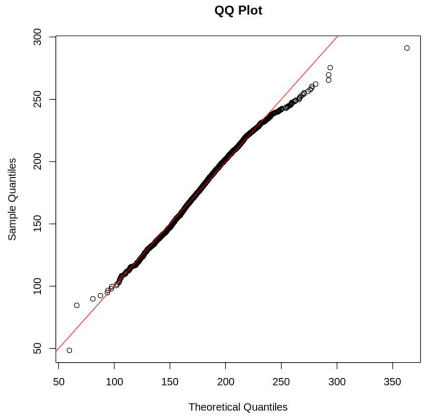
Zero Mean, Normality and Heteroscedasticity

Zero Mean:

- The label data is not zero mean . We shift the label data so as to.

Normality:

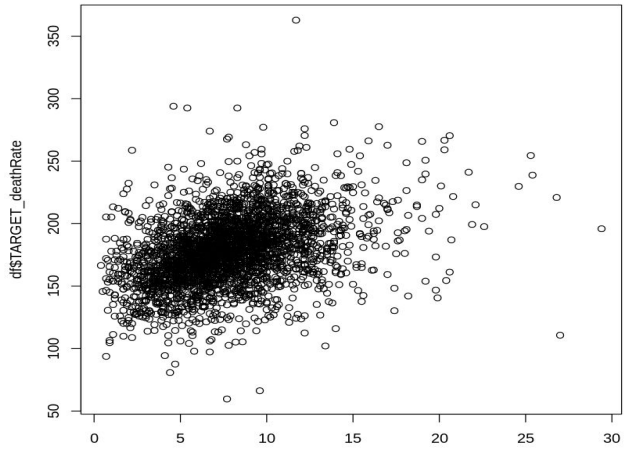


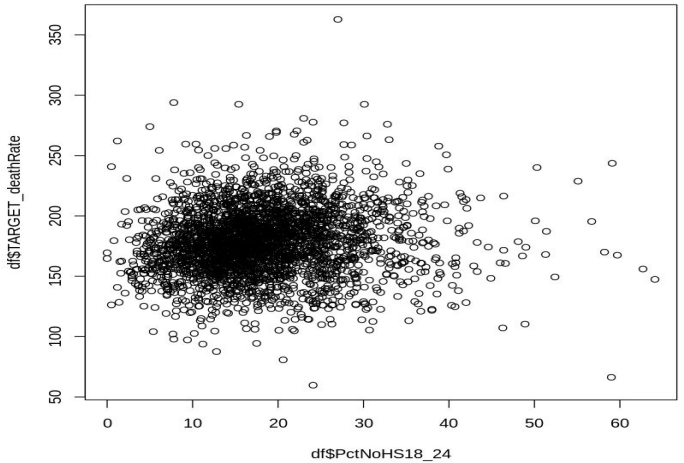


- *Symptoms*:- The Histogram of target variable seems Normal distributed.
- *Diagnostics*:- QQ line is very near to y=x line. Shapiro test yields a very small p-value,suggesting target variable is non normal.
- *Remedy*:- Box-cox Transformation suggested Square root transformation.

Heteroscedasticity :

- In our analysis we found a significant diversity across the distribution of young adults (18-24 years old) with low educational attainment (highest education below high school) and the unemployed population.





- *Diagnostics*:-*On performing Breush-pagan* test we found p-values smaller enough than 0.05(level of significance).
- *Remedy* :- After log Transformation, *Breush-pagan* test yields p-values larger than 0.05 .

Mediators and confounders: Distorts the true relationship(Y,X)

Cofounders:

- Unemployed people and Employed people are confounding variables for the cancer death rate . Thus we can remove one of two and keep the percentage of employed population out of total working population(age 16 over).
- Schooling features:- PctNoHS18_24,PctHS18_24,PctSomeCol18_24,PctBachDeg18_24,

  PctHS25_Over,PctBachDeg25_Over create confounding problem .

  With step function and a bit analysis we dropped all except PctBachDeg25_Over.

Mediators:

- Median female age and Median male age affects the cancer death rate through median age(Mediator).

Multicollinearity: Tests

Variance Inflation Factor (VIF) Analysis:

- Through a VIF analysis, we were able to identify predictors with high VIF values (greater than 10), signaling potential multicollinearity issues.
- Few of these predictors include, *avgDeathsPerYear* with a VIF value of 31.43, *popEst2015* with a VIF values of 26.42, and others.

Correlation Analysis:

- Correlation Matrix of the predictors were analyzed to identify the predictors with the coefficients near to 1.
- The correlation coefficients for the predictors the predictors *avgDeathsPerYear* and *avgAnnCount* is also high with the value 0.94, and for *PctBlack* and *PctWhite* variables is approximately -0.83.
- We find the correlation coefficient of the predictors is highest for *avgDeathsPerYear* and *popEst2015* to be 0.98 (almost 1).

Multicollinearity: Remedies

Predictor Combination and **dropping some features**:

- Here in this case as a remedy, we take the two predictor variables with the highest VIF values, *avgDeathsPerYear* and *popEst2015*, and create a new variable, *avgDeathsPerYearPercent* by transforming them into a percent ratio.

$$avgDeathsPerYearPercent = \frac{avgDeathsPerYear}{popEst2015} \times 100$$

- Dropped *PctBlack* and *avgDeathsPerYear* feature.

Principal Component Analysis (PCA):

- PCA was applied to the predictor variables to transform them into a set of linearly uncorrelated principal components, hence mitigating the issue of multicollinearity in the regression model.
- While PCA reduces multicollinearity, it also reduces the interpretability of the model significantly.

Whats next?

The core of the project.....

Regression

1. Model Description
2. Outliers
3. Analysis
4. Conclusion

Step function in R suggested features:

We employed a stepwise regression procedure to identify the most relevant features associated with cancer death rates. While acknowledging that stepwise selection might not yield the absolute best model, it provided valuable insights into key factors.

Stepwise regression identified significant factors using linear, quadratic, cubic, and logarithmic transformations of the independent variables.

Equation of Regression

Based on analysis done above , we arrived the following equations involving some linear and some logarithmic relations and Y is mean shifted log(cancer_death_rate) :-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{19} X_{19}$$

Y = log(cancer deaths per capita(1 lac)) , has mean 0.

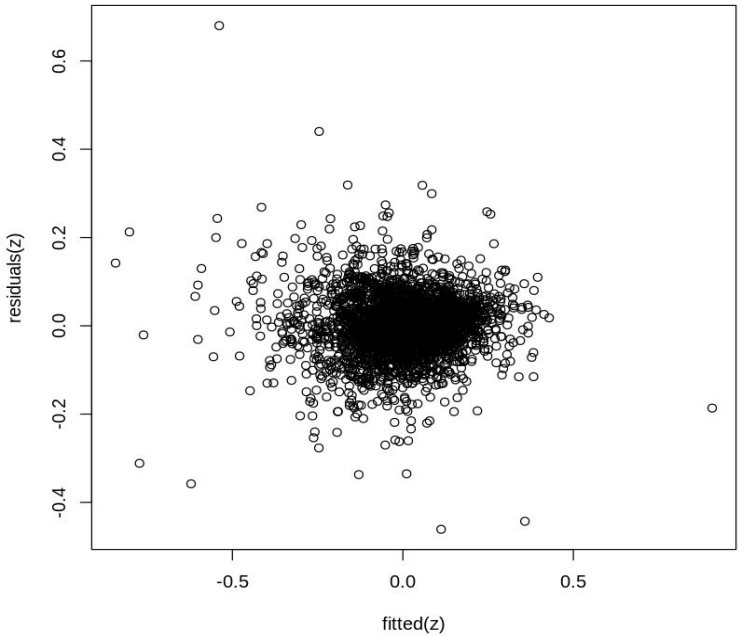
X1 = Mean of cancer cases annually, X2 = log(avgDeathsPerYear%) , X3 = incidence_rate ,X4 =MedianAgeFemale , X5 = MedianAgeMale , X6 = PctPublicCoverageAlone , X7 = Birthrate ,X8 = log(Poverty% , X9 = log(Studypercap), X10 = log(avghouseholdsize) , X11 = log(%married, x12= pctbachdegree25over), x13=log(pctunemplyed16over/(pctemplyed16over+pctunemplyed16over)), x14=log(pctprivatecoverage), x15= log(pctpubliccoverage), x16= log(pctwhite), x17=log(pctmarriedhousehold), x18=pctpubliccoveragealone,x19=log(medincome)

We got the accuracy from this model as **80.25**%.

Analysis :

1. The residuals plot shows no discernible patterns, suggesting that the model assumptions regarding the error terms is met. This indicates a good fit of the regression model.

2.The residuals are scattered relatively around zero, indicating low magnitudes of the residuals.



Effect of Outliers on the Model accuracy:

- The outliers were removed by employed Cook's distance method to identify influential outliers in the dataset.
- Following the removal of influential outliers, a notable improvement in the R-squared value of the regression model is observed.
- The increase in R-squared from approximately 0.80 to nearly 0.87 suggests that eliminating outliers enhanced the model's ability to explain the variance in the dependent variable.

Regression Model	Data Size	Adjusted R-Squared	Multiple R-Squared	Residual SE
With Outliers	3020	0.8025	0.8042	0.07065
After Outliers Removed	2807	0.8727	0.8715	0.05249

Closing Note

From analysing the regression model computed, we can conclude a few key points regarding the data:

- We notice that regions with higher Median income tend to have very low Target Death Rates related to Cancer.
- Another point that can be considered is from the predictor related to the Percentage of the population using Public (government-provided) Health Coverage, *PctPublicCoverage*. It is inversely related, that is, regions where a large ratio of the population with public health coverage tend to have a lesser Cancer Death Rates.
- Population of different races in the region, given by the variables such as *PctAsian*, have very low coefficients, showing that they do not have much influence on the Cancer Mortality Rates.

THANK YOU